ieve that those
ally (or, better
ases). Further-
for that matter)
fixed by perfu-

e provided with
ing the reasons
iation might be
ited chemicals.
one in the best
sions are often
system tissues.
ised in selected
or to detect a
ssays or immu-
dendritic arbor-
istructural eval-
information in
e applied in all
sults, it is time
neurotoxicolo-
ion is available
ming the in-life
gnment, the pa-
reach a false-
scopic sections
, these sections
about the possi-

histotechnology

:39461–39463.
Register. 1988 Feb

l Domestic Animals
-123, March 1991.
entanedione: 9-day
986; 7:329–339.
d hypermetabolism,

Neurobehavioral Toxicity: Analysis and Interpretation,
edited by B. Weiss and J. O'Donoghue.
Raven Press, Ltd., New York © 1994.

11

Statistical Issues for Animal Studies of Developmental Neurotoxicity

Christopher Cox

*Department of Biostatistics, University of Rochester Medical Center,
Rochester, New York 14642*

As with any area of scientific inquiry, and perhaps more than some, developmental neurotoxicology has unique issues of study design and data analysis, which are defined by the nature of both the questions posed and the resulting investigations. The term *developmental* usually refers to the period of life from conception to early adulthood. The focus on this period reflects the assumption that this is the most sensitive stage of life, both in terms of sensitivity to toxic insult and in terms of the magnitude of effects. This view is not universally held, as it has been pointed out that changes of a developmental nature may occur throughout life, or at least that the effects of exposure during the developmental period may not be apparent until much later in life. In fact, it has been argued that late life may represent a uniquely sensitive window for the observation of the effects of early exposure (1).

Within the early developmental period, the prenatal stage is generally considered to be most sensitive to toxic insult. Thus neither direct exposure nor direct observation are possible during this most sensitive period. Rather, the route of exposure must be through maternal administration, making the assessment of the actual exposure to the fetus problematic. In addition, measurements of behavioral effects must be made on the offspring after birth. Often linked to the premise of prenatal sensitivity is the assumption that the developing nervous system also displays a unique degree of sensitivity. This assumption leads to the expectation that effects of prenatal exposure should be most apparent in complex, developmentally appropriate tasks, and to the use of tests that reflect the capacities of the developing organism. In human subjects, such measures may include tests of various aspects of both motor and mental development. The Bayley Scales of Infant Development are a prototype. In animals, behavioral tests have ranged from examination of reflexes to the performance of complex, learned behaviors. The previous assumptions provide the context for the present discussion, which represents an expansion of a number of the core questions for this segment of the workshop.

18867

EXPERIMENTAL DESIGN

Purpose of the Study

It is important to emphasize at the outset that, in general, the answers to questions of study design and data analysis depend fundamentally on the purpose of the study. In developmental neurotoxicology, the most important distinction is between studies whose purpose is primarily hazard identification, especially studies designed to screen for potentially adverse effects, and those whose purpose is the investigation of a particular mechanism of toxicity, or the development of precise dose-response information. One consequence of the broad-gauge search for effects in the developing nervous system is the corresponding lack of information about mechanisms of action. Nevertheless, such studies are needed to place wholly empirical investigations on a more substantial scientific foundation. The mechanistic type of study will usually be much more focused than the hazard identification study because it limits the number of hypotheses to be tested. I would argue that while it may be necessary to examine large numbers of effects for some purposes, one cannot expect the resulting information to be as precise as that from a study that asks more specific questions and makes more focused measurements. I would also argue that although statistical methods may provide techniques for adjusting for a multiplicity of statistical analyses, as well as arriving at estimates of risk using data from such screening studies, statistical analysis cannot alleviate the basic problem, which is scientific, not statistical in nature.

This same distinction is helpful in dealing with questions of maternal toxicity. From the point of view of hazard identification, it may not be necessary to distinguish damage caused by a toxicant that acts directly on fetal tissue from damage to the mother, which in turn results in further harm to the fetus (2). Indeed, from a biological point of view, there may not be a sharp distinction, as in the case of metabolites produced in the mother, which subsequently cause damage to the fetus. If the effects of maternal damage are of interest for mechanistic understanding, then these could be studied separately from effects in the offspring. Furthermore, with many agents, fetal damage often occurs at exposure levels low enough that minimal maternal toxicity is observed. For the remainder of this discussion, it is assumed that the purpose of the study is the examination of developmental toxicity.

Choice of a Biological Test System

With regard to issues of statistical design, the first consideration is the choice of a test system. Perhaps the most important point to make is that screening studies may require a standardized test system, whereas a mechanistic study aims to achieve the best model for studying the mechanism of toxicity. Clearly, these different goals may result in the study of different aspects of toxicity of the same test substance. An example is the study of the effects of heavy metals, both in vivo and in vitro. In vitro test systems offer the potential for greater specificity in the study of mechanisms, although less sensitivity for screening a broad range of effects. In particular, such test systems cannot be used to test for behavioral effects, which must be observed on the whole animal as in a classic bioassay. The advantage of a standardized test system is that data from different studies are comparable to a greater extent; for example, it is easier to compare the relative toxicities of different test substances. The disadvantage is that the standard system may not be the most useful for examining any particular class of toxic effects.

Exposure Protocol

The next important question is one of exposure, and here again there are numerous possibilities. In observational studies in exposed human populations, exposure levels cannot be controlled, so that the discussion must be confined to studies in animals. In terms of the life cycle of the animal, the first possibility is prenatal exposure, for all or only part of gestation. If the gestational period can be naturally divided into different developmental stages, it might be informative for the study of mechanisms to have different groups of animals exposed at different gestational ages. This is a common strategy in teratology. Exposures early in gestation, however, may continue to produce effects throughout the entire gestational period by altering or disrupting the developmental sequence.

Postnatal exposure must also be considered for two different reasons. The first is that in some animal models, such as rodents, early postnatal development parallels late prenatal development in humans. Second, studies of lead exposure in human populations indicate that postnatal exposure levels predict later outcome better than prenatal, or at least maternal, levels. The postnatal period itself might also be divided into developmental stages. If both prenatal and postnatal exposures are specifically used, then it is not entirely clear which exposure values or combination of values should be used to predict outcome. Indeed, the optimal measure of exposure may depend on the substance under study, or on the outcome; in human studies, this is frequently the case. These reasons suggest that, as animals mature, either cumulative or recent exposure may be a better predictor of outcome than, for example prenatal exposure.

One recent study of lead exposure in children created a cumulative index of exposure by combining a number of measured blood lead concentrations (3). This index was then correlated with IQ scores. Such choices are easier if specific outcomes are of interest. It is essential that different levels of exposure be included, even in epidemiologic studies, if dose-effect relationships are to be obtained. A reasonable basis for the selection of experimental doses is toxicity to the mother (1). An alternative approach would be to choose exposures which model those in human populations. These might typically be chronic, low level exposures, at least if environmental as opposed to occupational exposures are of interest.

Measurement of Effect

The most numerous and, therefore, most difficult choices are those of the nature and timing of the outcome measurements. Because of the complexity of the nervous system, there is a wide array of possible testing strategies, and any particular test may lack either sensitivity to or specificity for a given test substance. In humans, standardized tests are available for assessing different domains of development. For animal studies, many different behavioral tests are available, and tests of reproductive function may in some cases be considered as well. For some purposes, a limited number of focused tests may be optimal. For other purposes, a standard battery covering the spectrum of developmental domains is preferable. Hypotheses about mechanisms of action favor selection of a more limited number of tests. A wide variety of different tests could provide insight into a range of possible mechanisms. In the absence of specific hypotheses, however, it may be difficult to connect behavioral test results with very concrete mechanisms of damage.

The timing of the various tests is also a critical issue. Different tests may be appropriate during particular windows of development, or it may be important to use certain tests longitudinally to monitor developmental processes that unfold over time. It has been argued that longitudinal follow-up is essential for developmental studies because the most subtle effects may appear as differences in rates over time, rather than measured levels of development at specified times. Alternatively, the best way to ascertain whether an effect is transient is to observe subjects longitudinally. Our ability to explicate trends, however, depends on our ability to describe them statistically (as e.g., linear vs. nonlinear) and on our ability to model the statistical dependencies introduced by repeated measurements.

An additional problem complicating the statistical properties of longitudinal data is that the analysis is more difficult if there are substantial numbers of missing values; this problem may be a consideration in deciding how long to follow a given group of animals. Another issue is the effects of repeated testing, although in many cases these effects are minimal. Situational adaptation and learning may introduce effects for which the experiment must control. In any case, it may be difficult to know the most sensitive developmental window for a particular end point, and repeated testing may be required for this reason.

A question receiving increasing attention is whether additional or unique effects are manifest later in life, and, in particular, whether early effects intensify or diminish with age. It has been suggested that the most subtle effects of early damage may occur at the end of life, as compensatory capacities diminish (1). In general, the optimum times for testing should depend on the presumed mechanisms or indices of damage. If there are no specific hypotheses, then testing would have to cover the entire life span. In contrast, if effects on biological substrates are of primary interest, then serial sacrifices may be required, which limits repeated testing of animals. The correlation of behavioral (whole animal) and biological (organs and tissues) measurements offers possibilities for the development of mechanistic hypotheses, which have not been fully explored.

Statistical Issues

An essential design consideration is the choice of an appropriate sample size, which is based on the magnitude(s) of the expected effect(s) and the desired statistical power. Sample size calculations may be difficult if a broad spectrum of effects is under study. Realistic sample size estimates may also be difficult if the statistical analysis is complex, for example, modeling longitudinal trends or interactions. In spite of such difficulties, some consideration of statistical power is essential before the study begins.

A second statistical design issue involves controlling for systematic effects, which could introduce bias. The scheduling of tests must be considered from a statistical as well as logistical perspective. As much as possible, testing schedules should be balanced with regard to factors such as time of day, day of the week, and testing apparatus, so that no bias is introduced. Such considerations also apply to caging and handling of animals. The use of standard experimental designs (4) should be considered, not only for the control of bias, but also because such designs provide increased precision for statistical comparisons. Spyker and Spyker (5) provide an example of a factorial design. Where possible formal randomization should be used.

ISSUES FOR STATISTICAL ANALYSIS

The foregoing questions are difficult—dealing with the design of developmental studies in neurotoxicology. The statistical analytic issues reflect these difficulties. The first of these deals with the nature of the individual end point(s) and the consequences for the corresponding statistical analysis. For statistical purposes, end points can be classified as either continuous (quantitative) or discrete (qualitative). In formulating a statistical analysis, the fundamental question is the nature of the statistical model, and different approaches are used for categorical than for continuous data. The family of Generalized Linear Models (6) provides a general framework for a large number of such models, and nonlinear models have been developed as well (7). For example, this family includes ordinary analysis of variance and regression models for continuous data, having uncorrelated, normally distributed errors, logistic regression models for binary (yes/no) data that have been widely used in epidemiologic studies, and classic, quantal (tumor/no tumor) response bioassays.

Many behavioral procedures provide quantitative as opposed to simply quantal data, leading to the expectation that ordinary regression-type methods will play a central role in the analysis of the resulting data. An example of a multiple linear regression approach to an unbalanced (unequal numbers in different groups), incomplete (not all combinations of factors) factorial design was described by Spyker and Spyker (5). The authors explored a linear model for main effects and selected interactions for the variables (factors) maternal dose, gestational day of administration, and prenatal versus postnatal (in either the biological or a foster mother) exposure. Consideration of interactions between dose level and other factors, such as day of gestation when the dose is administered, is an important part of the assessment of effects and a reflection of the complexities introduced by the different possible exposure schedules discussed above.

A second example of multiple regression analysis applied to developmental toxicity is found in Tachibana (8). He emphasized the use of the squared multiple correlation (R^2, or percentage of the total variation explained by the independent variables included in the regression model) as a measure of the usefulness of a given set of independent variables for explaining (variation in) the measure of outcome. A major advantage of ordinary regression models is that well-developed procedures exist for checking the required assumptions of normally distributed errors with constant variance.

It is helpful to think of a statistical model as having two components, a fixed part and a random part. The fixed part of the model usually includes the effects of the independent or predictor variables, such as dose, on the response. For example, in a multiple regression model, the fixed part of the model specifies that the mean response is a linear function of the independent variables, with unknown coefficients. The random part of the model specifies the distribution of the errors, the variation about regression. A major component of the error distribution of the statistical model is the dependency structure of the data. For many purposes, this structure can be summarized by the correlations among the observations on the dependent variable. In an ordinary regression model, the errors are assumed to be independent, so that no dependency is present. One source of statistical dependency is the serial dependency introduced by repeated measurements (with the same test) of the same animals and must be dealt with in the analysis of the data, for example, by performing a repeated measure as opposed to a conventional analysis of variance. A conceptually similar but statistically somewhat different source of dependency is the use of a number of different end points in a multivariate analysis. Here, the dependency structure is no longer serial with time, but is determined by the nature of the different tests, some being more highly correlated than others.

Another sort of dependency is shared by members of the same litter and is often referred to as the occurrence of litter effects. This question is relevant primarily when the offspring are exposed prenatally and must be considered in the analysis of the data, usually by taking the litter as the unit of analysis. Models for quantal responses that take account of litter size have been proposed (9,10). These models have the advantage of using responses from individual offspring without requiring the modeling of the dependency structure within each litter. Standard methods for continuous data are also available, but these require either assumptions about, or modeling of, dependencies over time or within each litter. For example, repeated measures analysis of variance has been used extensively. Although such analyses require strong assumptions about the dependency structure that are difficult to check and are probably frequently violated, adjustments to the significance tests have been developed that make these analyses much more robust (11). This kind of analysis, however, requires fairly complete data, and violations of important assumptions (for example, that the errors are normally distributed) can be difficult to detect. In addition, the assessment of effects often requires tests for interactions, and these are known to generally lack power compared with tests for main effects (12).

Extensions of repeated measures models to incomplete data have been developed, which also allow modeling of the dependency structure (13). The family of quasi-likelihood models provides a flexible framework for modeling data with dependencies, including correlated binary data (6). The structure of the statistical dependencies assumed by the model should reflect the testing situation, e.g., longitudinal testing or litter effects, and modeling the appropriate correlations will be important for the analysis.

Modeling the dependency structure is much more difficult than modeling of response levels (means or rates), and the effects of errors in such modeling are not well characterized, but can be substantial. For example, in the repeated measures context, use of the standard analysis of variance, with no adjustment to the significance tests for violation of assumptions, can lead to very different results than those for the adjusted analysis. The same argument applies in the case of multivariate analysis of a battery of tests. Although multivariate analysis may be helpful in some situations, each end point requires some sort of individual assessment. Although multivariate methods may provide an approach to the problem of multiplicity, it has already been argued that the real problem is multiplicity of hypotheses, not statistical analyses. A more promising use of multivariate methods is as tools for data reduction. For an example of this use with multiple outcome variables see Tepper et al. (14). For longitudinal data with a relatively large number of repeated measurements, polynomial regression has been proposed as one approach to summarizing trends in the data for individual animals (15). A low-order polynomial is fitted to data from each animal, and statistical comparisons of treatment groups are based on the fitted regression coefficients.

Design issues concerning studies involving litters of animals include how many offspring to select from each litter and whether both males and females should be included. It has been argued that the second point is important, even if sex differences are not of direct interest in the study. The standardization of litter sizes and sex distributions helps to produce more uniform data and is important even if the litter is the unit of analysis, because even with correlated responses, the precision of the average measurement should increase with the size of the litter. The advantage of such uniformity is that the analysis of balanced data (equal group sizes for all combinations of factors) is much simpler. For example, in the study of Spyker and Spyker (5) the full range of interactions could easily have been explored had the data been balanced. In addition, if a small number of animals from each litter is used, or if animals are exposed postnatally as well as prenatally, or tested later in life, then the magnitude of litter effects may be diminished because other effects intervene.

A related question concerns the choice of appropriate control groups. In some cases either paired feeding or cross-fostering must be considered, the latter as a method for separating the effects of prenatal and postnatal exposures. The most comprehensive approach to this question would involve four groups in a factorial arrangement, prenatal (yes/no) versus postnatal (yes/no) exposure (5).

The problem of missing data may be difficult to avoid. In particular, animals who die before the end of the study may be especially important, as they may be the most vulnerable or severely affected. When the greatest number of deaths occurs in the group receiving the highest dose, or when there is other evidence that the deaths were related to the exposure, there is a danger of underestimating the magnitude of the effect. For some purposes, it may be reasonable to omit or replace such animals; in other cases it may be possible to use incomplete data in the analysis. It is wise, however, not to assume that statistical methods can deal with all missing data problems. The choice of exposure levels in a long-term study deserves careful consideration, and the absence of acute effects may not be a reliable guide. A related question is whether the health status of animals must be monitored, either for quality control purposes, or because it may be important to include such data in the analysis.

In some studies, different tests are given to different groups of animals, and the problem is to combine results rather than combine data. In this case, the techniques of meta-analysis may be used, either to combine standardized measures of effect, or to combine *P* values. When many tests are used, the problem of the inflation of the type I (false-positive) error rate by multiple significance tests always lurks. It was argued earlier that this problem is inherent in any exploratory (hazard identification) study. The best solution to this problem is critical interpretation of the results, which is not inherently a statistical question.

One way to structure this interpretation is to divide the outcome variables into primary (confirmatory) and secondary (exploratory) categories. Primary outcomes are those used to address the specific hypotheses of interest. Another approach is to create groups of similar end points and to require that a reasonable consistency (consistency with regard to the nature, i.e., direction, of effects is more important than consistency with regard to statistical significance) be evident among similar outcomes. Multivariate statistical methods and the associated multiple comparison procedures offer an approach to the analysis of multiple end points, but substantially increase the burden of assumptions. In the absence of a limited series of primary hypotheses, a study should be considered exploratory, and it is important not to overinterpret the results of any exploratory study. Indeed, a good deal of restraint may be required in the case of a study which uses a battery consisting of a large number of tests.

A bound on the inflation of type I error is provided by the Bonferroni inequality, which states that if k tests are used, then the appropriate level of significance for a 5% overall error rate is 0.5/k. Thus, if a battery consists of 20 tests, each must be assessed at a significance level of 0.05/20 = 0.0025. This criterion is conservative, but it gives an idea of the size of the problem. Although a *P* value of 0.01 would be clearly nonsignificant by this criterion, the temptation might be strong to report a highly significant result. This is especially true in the climate of hazard identification, where the conservative philosophy of the scientific method, which is reflected

in the basic
null hypoth
further disc

I would li
comments o

1. Weiss B. R
ogy 1990;1
2. Tilson H.
1992;14:19
3. Baghurst P
intelligence
4. Cochran W
5. Spyker DA,
tal effects o
40:511–527
6. McCullagh
7. Cox C. Gen
8. Tachibana T
reanalysis o
icological R
9. Rai K, Van
sponses. *Bio*
10. Gaylor DW,
mental defec
11. Tamura RN,
studies. *Neur*
12. Fleiss JL. *Th*
13. Jennrich RI,
matrices. *Bio*
14. Tepper JL, V
Pharmacol 1
15. Cox C, Cory-
toxicology, *F*

in the basic framework of statistical hypothesis testing with its preference for the null hypothesis, gives way to the politics of risk assessment. For an example and further discussion see Tilson (2).

ACKNOWLEDGMENT

I would like to thank Drs. Bernard Weiss and Deborah Cory-Slechta for their comments on earlier drafts of this manuscript.

REFERENCES

1. Weiss B. Risk assessment: the insidious nature of neurotoxicity and the aging brain. *Neurotoxicology* 1990;11:305–314.
2. Tilson H. Study design considerations in developmental neurotoxicology. *Neurotoxicol Teratol* 1992;14:199–203.
3. Baghurst PA, McMichael AJ, Wigg NR, et al. Environmental exposure to lead and children's intelligence at the age of seven years. *N Engl J Med* 1992;327:1279–1284.
4. Cochran WG, Cox GM. *Experimental designs*. 2nd ed. New York: John Wiley; 1957.
5. Spyker DA, Spyker JM. Response model analysis for cross-fostering studies: prenatal versus postnatal effects on offspring exposed to methylmercury dicyandiamide. *Toxicol Appl Pharmacol* 1977; 40:511–527.
6. McCullagh P, Nelder JA. *Generalized linear models*. 2nd ed. New York: Chapman and Hall; 1989.
7. Cox C. Generalized linear models—the missing link. *Appl Statistics* 1984;33:18–24.
8. Tachibana T. Behavioral teratogenic insult of methylmercury assessed by using a set of measures: reanalysis of data from the Collaborative Behavioral Teratology Study of National Center for Toxicological Research, *Physiology and Behavior* 1989;45:1243–1247.
9. Rai K, Van Ryzin J. A dose-response model for teratological experiments involving quantal responses. *Biometrics* 1985;41:1–9.
10. Gaylor DW, Razzaghi M. Process of building biologically based dose-response models for developmental defects. *Teratology* 1992;46:573–581.
11. Tamura RN, Buelke-Sam J. The use of repeated measures analyses in developmental toxicology studies. *Neurotoxicol Teratol* 1992;14:205–210.
12. Fleiss JL. *The design and analysis of clinical experiments*. New York: John Wiley; 1986.
13. Jennrich RI, Schluchter MD. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* 1986;42:805–820.
14. Tepper JL, Weiss B, Cox C. Microanalysis of ozone depression of motor activity. *Toxicol Appl Pharmacol* 1982;64:317–326.
15. Cox C, Cory-Slechta DA. Analysis of longitudinal time series data from experiments in behavioral toxicology, *Fundam Appl Toxicol* 1987;8:159–169.

ires. The most
s in a factorial
).
r, animals who
ay be the most
s occurs in the
that the deaths
magnitude of
such animals;
sis. It is wise,
sing data prob-
eful considera-
elated question
quality control
analysis.
imals, and the
the techniques
es of effect, or
nflation of the
s lurks. It was
identification)
of the results,

variables into
ary outcomes
approach is to
le consistency
nore important
among similar
le comparison
t substantially
es of primary
portant not to
al of restraint
ing of a large

ni inequality,
nificance for a
each must be
conservative,
0.01 would be
ng to report a
ard identifica-
ch is reflected

Neurobehavioral Toxicity

Analysis and Interpretation

Editors

Bernard Weiss, Ph.D.

Professor
Department of Environmental Medicine
University of Rochester School of Medicine and Dentistry
University of Rochester Medical Center
Rochester, New York

John L. O'Donoghue, V.M.D., Ph.D.

Director
Corporate Health Environment Laboratories
Eastman Kodak Company
Rochester, New York

Raven Press **New York**

Raven Press, Ltd., 1185 Avenue of the Americas, New York, New York 10036

Made in the United States of America

Library of Congress Cataloging-in-Publication Data

Neurobehavioral Toxicity : analysis and interpretation / edited by
Bernard Weiss, John O'Donoghue.
p. cm.
Includes bibliographical references and index.
ISBN 0-7817-0172-4
1. Behavioral toxicology—Congresses. I. Weiss, Bernard.
II. O'Donoghue, John L.
RA1224.N493 1994
616.8—dc20 93-48577
CIP

The material contained in this volume was submitted as previously unpublished material, except in the instances in which credit has been given to the source from which some of the illustrative material was derived.

Great care has been taken to maintain the accuracy of the information contained in the volume. However, neither Raven Press, Ltd. nor the editors can be held responsible for errors or for any consequences arising from the use of the information contained herein.

Materials appearing in this book prepared by individuals as part of their official duties as U.S. Government employees are not covered by the above-mentioned copyright.

9 8 7 6 5 4 3 2 1

Ci